Arabic Semantic Similarity Approaches – Review

Marwah Alian
Hashemite University
Princess Sumaya University for Technology
Marwah2001@yahoo.com

Arafat Awajan
Princess Sumaya University for Technology
Amman, Jordan
awajan@psut.edu.jo

***Abstract*— Semantic similarity between different texts is a challenging task. It measures the relation between texts, sentences and words to depict their degree of similarity or resemblance. Many of the Natural languages processing tasks and applications need measuring the text semantic similarity to achieve good results. Plagiarism detection, text entailment, text summarization, machine translation, and information extraction are among these applications. Therefore many Methods for measuring semantic similarity have been proposed for Arabic text which is reviewed and compared in this research.**

Keywords—Semantic Similarity; Sentence Similarity; word Similarity; document Similarity; Neural Networks; word embeddings; Latent semantic analysis.

I. INTRODUCTION

The similarity between sentences or documents depends mainly on the similarity between words as they are the smallest component of documents or sentences. However, the similarity between words taken out from their context cannot give accurate results and should be measured according to their syntactical and or semantics features.

Measuring semantic similarity is considered an important part in various Natural Language Processing tasks and applications such as wordsense disambiguation, text summarization, entailment, machine translation and more [1]. Most of the existing word similarity measures are developed and used for English texts while very rare measures have been developed specifically for Arabic. The rare approaches used with Arabic languages are mainly adapted versions from those used for English texts. For example, the work of Almarsoomi et al. [2] represents an approach for measuring semantic similarity between two Arabic words which is originally depends on Li similarity measure [3]

In this work, we review the effort done by researchers for the task of measuring semantic similarity for Arabic text. We categorize existing researches into document similarity, sentence similarity and word similarity then we compare between these proposed approaches.

This research is organized as follows: section 2 define similarity concept, section 3 categorize and describe similarity approaches for Arabic text, then a comparison between existing approaches in section 4 and finally the summary is in section 5.

II. SIMILARITY DEFINITION

Similarity is a widely discussed concept in the field of linguistics, philosophy and information theory. Some of them categorize semantic relations according to the result from judgment of likeness or difference. For example synonyms, paraphrasing, and entailment are considered standard semantic relations while antonyms, inconsistency and contradiction resulted from the judgment of difference. However, two text units are considered to be similar if they have concentrated on a common action, concept, or object. Also this common object or actor should be the subject to the same action [4]. A universal definition of similarity in terms of information theory is presented in [5] where the measure of similarity is derived from a set of a assumptions while the universality of the similarity definition can be applied in different domains.

The similarity between two objects is related to their commonality and distance where the more commonality they share the more similar they are and the more difference they have the less similar they are. Also, the maximum similarity between two objects is reached when they are identical, no matter how much commonality they share. While semantic similarity is defined as a measure of conceptual distance between the two compared words or objects depending on the correspondence of their meaning [5].

In [6], the authors consider the semantic similarity as a challenging task if you are given two texts then the challenge is to measure how similar they are or to decide if they have a qualitative semantic relation between them such as paraphrase relation where two texts share the same meaning or entailment relation when one text is inferred from the other.

Semantic similarity is defined as a confidence score which reflects the semantic relation between two short texts (sentences) where the higher the score the more similar meaning the two texts have [7]. According to Li the meaning of a sentence is reflected by the words constructing the sentence.

III. SEMANTIC SIMILARITY FOR ARABIC TEXT

A. Document Similarity

Soori et al. [8] use the similarity between documents in plagiarism detection for Arabic text.

978-1-7281-0385-3/18/$31.00 ©2018 IEEE